(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

Übersetzung der europäischen Patentschrift

(97) EP 0 573 301 B1

(10) DE 693 24 629 T 2

(51) Int. Cl.[6]: G 10 L 5/06
G 10 L 7/08

(21) Deutsches Aktenzeichen: 693 24 629.4
(96) Europäisches Aktenzeichen: 93 304 340.8
(96) Europäischer Anmeldetag: 4. 6. 93
(97) Erstveröffentlichung durch das EPA: 8. 12. 93
(97) Veröffentlichungstag der Patenterteilung beim EPA: 28. 4. 99
(47) Veröffentlichungstag im Patentblatt: 30. 9. 99

(30) Unionspriorität:
922606 05. 06. 92 FI

(73) Patentinhaber:
Nokia Mobile Phones Ltd., Salo, FI

(74) Vertreter:
TER MEER STEINMEISTER & Partner GbR
Patentanwälte, 81679 München

(72) Erfinder:
Ranta, Jukka Tapio, SF-24130 Salo, FI

(84) Benannte Vertragstaaten:
DE, FR, GB, SE

(54) Verfahren und Vorrichtung zur Spracherkennung

DE 693 24 629 T 2

DE 693 24 629 T 2

Europäisches Patent
mit der Europäischen Patentanmeldungsnr. 93 304 340.8
NOKIA MOBILE PHONES LTD.
Case: PAT 92518

Verfahren und Vorrichtung zur Spracherkennung

Die Erfindung betrifft ein Verfahren und eine Vorrichtung zur Spracherkennung, insbesondere ein Verfahren und ein System für ein durch Sprache steuerbares Telefon, wobei ein Wert eines Bezugsworts durch eine Spracherkennungseinrichtung auf Grundlage eines von einem Benutzer gesprochenen Worts berechnet wird und eine Erkennungsauflösung auf Grundlage dieses Werts erstellt wird.

Telefone sind im Allgemeinen mit einem Handapparat versehen, den der Benutzer in der Hand hält, während er spricht. Dies gilt auch für den Fall, dass Funktelefone wie Mobiltelefone verwendet werden. Bei einem derartigen Telefon bleibt nur eine Hand frei, was beim Fahren zu Schwierigkeiten führen kann. Eine Lösung hinsichtlich dieses Problems besteht in einem im Auto angebrachten gesonderten Mikrophon sowie einem gesonderten Lautsprecher, der auf eine geeignete Lautstärke einzustellen ist und mit geeignetem Abstand vom Benutzer positioniert ist, so dass der Benutzer den anderen Teilnehmer deutlich hören kann. Selbst bei diesem Design muss der Benutzer eine seiner Hände verwenden, um einen Anruf zu tätigen, d.h. zum Wählen der Nummer der anderen Partei oder zum Reagieren auf einen eingehenden Anruf oder zum Beenden eines Anrufs.

Damit sich ein Telefonbenutzer auf das Fahren konzentrieren kann, wurden sogenannte Freisprechtelefone entwickelt, bei denen die Funktionen durch Sprache steuerbar sind. Hierbei können alle Telefonfunktionen durch Sprache gesteuert werden, wie das Ein/Ausschalten, Senden/Empfangen, Sprachlautstärke-Steuerung, Wählen einer Telefonnummer, Antworten auf einen Telefonanruf, und so kann sich der Benutzer auf das Fahren konzentrieren. Der Fahrer muss seine Hände nicht vom Lenkrad wegnehmen und seine Augen nicht von der Straße ablenken, weswegen ein Freisprechtelefon die Fahrsicherheit beträchtlich erhöht.

Ein Nachteil in Zusammenhang mit einem sprachgesteuerten Telefon besteht darin, dass die Spracherkennung nicht völlig perfekt ist. Durch die Fahrzeugumgebung hervorgerufene Hintergrundgeräusche sind stark, weswegen die Spracherkennung schwieriger ist. Es erfolgten etliche Anstrengungen zum

Vermarkten der Spracherkennungsfähigkeit in Zusammenhang mit Mobiltelefonen, jedoch war angesichts der Unzuverlässigkeit von sprachgesteuerten Telefonen das Interesse von Benutzern an solchen unbedeutend. Die Erkennungsgenauigkeit von in der Technik bekannten Spracherkennungseinrichtungen ist nicht sehr gut, insbesondere unter ungünstigen Bedingungen, z. B. in einem fahrenden Fahrzeug, in dem die starken Hintergrundgeräusche eine zuverlässige Worterkennung im Wesentlichen verhindern. Fehlerhafte Erkennungsauflösungen verursachen im Allgemeinen die größten Unbequemlichkeiten beim Realisieren eines Benutzer-Kommunikationssystems, da sie unerwünschte Funktionen starten können, wie die Beendigung von Anrufen in deren Verlauf, was aus dem Gesichtspunkt des Benutzers besonders unzweckdienlich ist. Die üblichsten Konsequenzen fehlerhaften Sprachinterpretierung bestehen im Wählen einer falschen Nummer. Aus diesem Grund sind Benutzerkommunikationen so konzipiert, dass durch eine Spracherkennungseinrichtung keinerlei Erkennungsauflösung erfolgt, wenn keine ausreichende Sicherheit hinsichtlich eines vom Benutzer gesprochenen Worts erzielt ist, wobei in derartigen Fällen der Benutzer im Allgemeinen dazu aufgefordert wird, den geäußerten Befehl zu wiederholen.

Nahezu alle Spracherkennungseinrichtungen beruhen auf dem Funktionsprinzip, dass ein von einem Benutzer gesprochenes Wort durch ein ziemlich kompliziertes Verfahren mit zuvor in den Speicher der Spracherkennungseinrichtung eingespeicherten Bezugswörtern verglichen wird. Spracherkennungseinrichtungen berechnen im Allgemeinen eine jedem Bezugswort entsprechende Zahl, die anzeigt, in welchem Ausmaß das vom Benutzer gesprochene Wort dem Bezugswort ähnelt. Abschließend erfolgt eine Erkennungsauflösung auf Grundlage der Zahlen in solcher Weise, dass für die Auflösung dasjenige Bezugswort gewählt wird, dem das geäußerte Wort am meisten ähnelt. Eines der bekanntesten Verfahren für den Vergleich zwischen einem gesprochenen Wort und den Bezugswörtern ist das Dynamic-Time-Warping(DTW)-Verfahren und das statistische Hidden-Markov-Modell(HMM)-Verfahren.

Sowohl beim DTW- als auch beim HMM-Verfahren wird ein unvertrautes Sprachmuster mit den bekannten Bezugsmustern verglichen. Beim Dynamic-Time-Warping wird ein Sprachmuster in eine Anzahl von Rahmen unterteilt, und es wird der örtliche Abstand zwischen dem Sprachteil in jedem Rahmen und dem dem Bezugsmuster entsprechenden Sprachteil berechnet. Auf Grundlage der auf diese Weise hergeleiteten örtlichen Abstände wird durch einen DTW-Algorithmus nach dem minimalen Pfad zwischen dem Anfangs- und dem Endpunkt des Worts gesucht. So kann durch Dynamic-Time-Warping ein Abstand zwischen dem

gesprochenen Wort und den Bezugswörtern erhalten werden. Beim HMM-Verfahren werden Sprachmuster erzeugt, und dieser Sprachmuster-Erzeugungsschritt wird durch ein Statusänderungsmuster gemäß dem Markov-Verfahren strukturiert. Dieses Statusänderungsmuster ist so das HMM. Spracherkennung für die empfangenen Sprachmuster erfolgt nun durch Definieren der Beobachtungswahrscheinlichkeit für diese Sprachmuster unter Zuhilfenahme des HMM-Musters. Unter Verwendung des HMM bei der Spracherkennung wird als erstes ein HMM-Muster für jedes zu erkennende Wort, d.h. für jedes Bezugswort, erzeugt. Die HMM-Muster werden in den Speicher der Spracherkennungseinrichtung eingespeichert. Nachdem die Spracherkennungseinrichtung das Sprachmuster empfangen hat, wird für jedes im Speicher gespeicherte HMM-Muster eine Beobachtungswahrscheinlichkeit berechnet, und im Ergebnis des Erkennungsprozesses wird ein Wort für dasjenige HMM-Muster geliefert, für das die höchste Beobachtungswahrscheinlichkeit erhalten wurde. Anders gesagt, wird für jedes Bezugswort die Wahrscheinlichkeit berechnet, gemäß der es das vom Benutzer gesprochene Wort wäre. Die oben genannte höchste Beobachtungswahrscheinlichkeit beschreibt die Gleichheit des empfangenen Sprachmusters mit dem nächstkommenden HMM-Muster, d.h. dem nächstkommenden Bezugssprachmuster.

So berechnet die Spracherkennungseinrichtung bei den aktuellen Systemen eine bestimmte Zahl für die Bezugswörter auf Grundlage des von einem Benutzer gesprochenen Worts; beim DTW-System ist die Nummer der Abstand zwischen Wörtern, und beim HMM-Verfahren zeigt die Nummer die Wahrscheinlichkeit der Gleichheit der Wörter an. Wenn das HMM-Verfahren verwendet wird, wird im Allgemeinen für die Spracherkennungseinrichtungen eine vorgegebenen Schwellenwahrscheinlichkeit definiert, die das wahrscheinlichste Bezugswort erreichen muss, um eine Erkennungsauflösung zu liefern. Ein anderer Faktor, der die Erkennungsauflösung beeinflusst, könnte z. B. die Differenz zwischen den Wahrscheinlichkeiten für das wahrscheinlichste Wort und das zweitwahrscheinlichste Wort sein; es ist zu erwarten, dass sie ausreichend groß ist, damit eine Erkennungsauflösung erfolgen kann. Wenn eine Erkennungsauflösung auf Grundlage der Erkennungswahrscheinlichkeit für das wahrscheinlichste Wort erfolgt, soll die Irrungswahrscheinlichkeit höchstens z. B. 0,1 betragen. Daher ist es möglich, dass dann, wenn Hintergrundgeräusche stark sind, für ein Bezugswort im Speicher, wie das Bezugswort "1", auf Grundlage eines vom Benutzer geäußerten Befehls bei jedem Versuch z. B. 0,8 als größte Wahrscheinlichkeit beim Vergleich mit den anderen Bezugswörtern erhalten wird. Da die Wahrscheinlichkeit unter der Schwellenwahrscheinlichkeit von 0,9 bleibt, wird das Wort nicht akzeptiert und es kann erforder-

lich sein, dass der Benutzer den Befehl mehrfach äußern muss, bevor die Grenze der Erkennungswahrscheinlichkeit überschritten wird und die Spracherkennungseinrichtung den Befehl akzeptiert, obwohl die Wahrscheinlichkeit sehr dicht am Akzeptierwert gelegen haben kann. Vom Gesichtspunkt des Benutzers her ist dies höchst störend. Ein korrektes Erkennungsergebnis kann beim ersten Versuch unter Verwendung der aktuellen Technik ziemlich häufig dann erzielt werden, wenn die Geschwindigkeit eines Fahrzeugs unter 80 bis 90 km pro Stunde liegt, abhängig von der Geräuschisolierung des Wagens und der Sprechweise des Benutzers. Bei höheren Geschwindigkeiten nimmt jedoch die Funktion der Erkennungseinrichtung sehr stark ab, und in den meisten Fahrzeugen arbeitet die Spracherkennungseinrichtung bei Geschwindigkeiten über 100 km pro Stunde nicht mehr ausreichend zuverlässig dafür, dass sie als nützlich angesehen werden könnte. Insbesondere bei derartigen Geschwindigkeiten ist aber das Erfordernis, die Verkehrssicherheit zu erhöhen, größer als bei niedrigeren Geschwindigkeiten.

Das US-Patent Nr. 4 783 803 offenbart ein Spracherkennungssystem, das die akustische Ähnlichkeit zwischen einem gesprochenen Wort und einem Bezugswort sowie eine Sprachmodellbewertung auf Grundlage zuvor erkannter Wörter kombiniert. Ein derartiges bekanntes System nutzt die in einem Sprachmodell enthaltene A-priori-Wahrscheinlichkeit, dass ein gegebenes Wort vom Benutzer gesprochen wird, wenn ein zuvor erkanntes Wort oder mehrere vorgegeben sind. Das US-Patent Nr. 5 003 603 offenbart ein Spracherkennungssystem, bei dem dann, wenn ein gesprochenes Wort gemäß Vergleichskriterien nicht mittels irgendeines Bezugsworts erkannt werden kann, das System den Benutzer dazu auffordert, die Äußerung zu wiederholen.

Gemäß einer ersten Erscheinungsform der Erfindung ist eine Spracherkennungsvorrichtung mit folgendem geschaffen: einer Vergleichseinrichtung zum Vergleichen eines von einem Benutzer gesprochenen ersten Worts mit mindestens einem vorbestimmten Bezugswort; einer Berechnungseinrichtung zum Berechnen eines Werts, der der Ähnlichkeit zwischen dem vom Benutzer gesprochenen ersten Wort und dem mindestens einen vorbestimmten Bezugswort entspricht; einer Auswähleinrichtung zum Auswählen des Werts, der der größten Wahrscheinlichkeit entspricht; dadurch gekennzeichnet, dass die Berechnungseinrichtung so ausgebildet ist, dass sie den ausgewählten Wert beim Berechnen eines neuen Werts entsprechend der Ähnlichkeit zwischen einem zweiten vom Benutzer gesprochenen Wort und dem mindestens einen Bezugswort verwendet, wenn der ausgewählte Wert einem vorbestimmten Kriterium genügt.

Gemäß einer zweiten Erscheinungsform der Erfindung ist ein Spracherkennungsverfahren geschaffen, das folgendes umfasst: Vergleichen eines von einem Benutzer gesprochenen ersten Worts mit mindestens einem vorbestimmten Bezugswort; Berechnen eines Werts, der der Ähnlichkeit zwischen dem vom Benutzer gesprochenen ersten Wort und dem mindestens einen vorbestimmten Bezugswort entspricht; Auswählen des Werts, der der größten Ähnlichkeit entspricht; dadurch gekennzeichnet, dass der ausgewählte Wert dazu verwendet wird, einen neuen Wert entsprechend der Ähnlichkeit zwischen einem vom Benutzer gesprochenen zweiten Wort und dem mindestens einen Bezugswort zu berechnen, wenn der ausgewählte Wert einem vorbestimmten Kriterium genügt.

Die Erfindung hat den Vorteil, dass eine zuverlässigere Erkennung von Wörtern selbst dann möglich ist, wenn die Ähnlichkeit zwischen gesprochenen Wörtern und Bezugswörtern nicht hoch ist.

Bei einer Ausführungsform gemäß der ersten und zweiten Erscheinungsform der Erfindung ist mehr als ein Bezugswort vorhanden. Dies hat den Vorteil, dass der Benutzer bei einer die Erfindung enthaltenden Steuerungsvorrichtung, die über mehr als eine sprachgesteuerte Funktion verfügt, Sprachsteuerung verwenden kann.

Bei einer bevorzugten Ausführungsform der ersten und zweiten Erscheinungsform der Erfindung ist das vom Benutzer gesprochene zweite Wort dasselbe wie das von ihm gesprochene erste Wort. Dies hat den Vorteil, dass eine zweite Berechnung nur dann ausgeführt wird, wenn das zweite gesprochene Wort mit dem ersten gesprochenen Wort übereinstimmt, um dadurch eine unnötige Verzögerung bei der Erkennung gesprochener Wörter zu vermeiden.

Bei einem alternativen Ausführungsbeispiel der ersten und zweiten Ausführungsform der Erfindung wird der ausgewählte Wert nur dann beim Berechnen eines neuen Werts verwendet, wenn das vom Benutzer gesprochene zweite Wort dasselbe wie das vom ihm gesprochene erste Wort ist. Dies hat den Vorteil, dass unnötige Berechnungen vermieden sind und dass ein voriger Wert nur dazu verwendet wird, die Erkennung eines wiederholt vom Benutzer gesprochenen Worts zu unterstützen.

Vorzugsweise besteht das vorbestimmte Kriterium darin, dass der ausgewählte Wert kleiner als ein vorbestimmter Schwellenwert ist oder alternativ das vorbestimmte Kriterium darin besteht, dass die Differenz zwischen dem ausgewählten Wert und einem anderen Wert, der der Ähnlichkeit zwischen dem

ersten vom Benutzer gesprochenen Wort und einem anderen Bezugswort entspricht, kleiner als ein vorbestimmter Schwellenwert ist. Dies hat den Vorteil, dass weitere Äußerungen und Berechnungen nur dann erforderlich sind, wenn ein gesprochenes Wort nicht zuverlässig erkannt werden kann oder wenn ein gesprochenes Wort zwei verschiedenen Bezugswörtern ähnlich ist.

In geeigneter Weise wird eine Wiederholung des vom Benutzer gesprochenen ersten Worts dann angefordert, wenn der ausgewählte Wert das vorbestimmte Kriterium erfüllt, was dem Benutzer deutlich anzeigt, dass ein gesprochenes Wort nicht erkannt wurde und dass er das Wort wiederholen muss.

Beim erfindungsgemäßen Verfahren berechnet eine Spracherkennungseinrichtung die Erkennungswahrscheinlichkeiten für Bezugswörter, und sie erzeugt eine Erkennungsauflösung, wenn eine der Wahrscheinlichkeiten einen vorbestimmten Schwellenwert überschreitet; andernfalls wird der Benutzer dazu aufgefordert, das Wort erneut zu sprechen, und dafür erfolgt eine Erkennungsauflösung, wenn die Wahrscheinlichkeit für eines der Bezugswörter kleiner als ein vorbestimmter Schwellenwert ist; andernfalls wird eine neue Wahrscheinlichkeit unter Verwendung der von der Spracherkennungseinrichtung berechneten aktuellen Wahrscheinlichkeit und einer Wahrscheinlichkeit, die einmal oder mehrere Male zuvor berechnet wurde, unter der Bedingung berechnet, dass es sich um Wahrscheinlichkeiten über ein und dasselbe Bezugswort handelt, wobei eine Erkennungsauflösung dann erzeugt wird, wenn die Wahrscheinlichkeit einen vorbestimmten Schwellenwert überschreitet. Solange der vorbestimmte Schwellenwert nicht durch die durch die Spracherkennungseinrichtung berechnete Wahrscheinlichkeit überschritten ist, wird die berechnete Wahrscheinlichkeit in den Speicher eingespeichert, der Benutzer wird dazu aufgefordert, das Wort erneut zu sprechen, und der im Speicher gespeicherte Wert wird zusammen mit der folgenden Wahrscheinlichkeit / den folgenden Wahrscheinlichkeiten verwendet, wie sie für dasselbe Wort von der Spracherkennungseinrichtung berechnet wurden, um eine neue Wahrscheinlichkeit zu berechnen, die auf Grundlage der Wahrscheinlichkeiten zu berechnen ist (um eine Erkennungsauflösung zu erzeugen, wenn, unter Berücksichtigung der vorangehenden Wahrscheinlichkeiten, die Schwellenwahrscheinlichkeit erreicht ist). Danach wird, wenn die Spracherkennungseinrichtung eine den Schwellenwert überschreitende Wahrscheinlichkeit berechnet, oder dieser unter Berücksichtigung der vorangehenden Wahrscheinlichkeiten erreicht wird, der Speicher rückgesetzt. Auch dann, wenn eine Wiederholung eines vorigen Worts fraglich ist, wird der Speicher vor einer Erkennungsauflösung rückgesetzt. Der Speicher wird auch dann rückgesetzt, wenn die Spannung in

der Vorrichtung eingeschaltet wird und wenn ein Vorgang unterbrochen wird.

Die Erfindung wird unten nur beispielhaft und unter Bezugnahme auf die beigefügten Zeichnungen im einzelnen beschrieben.

Fig. 1 zeigt ein Prinzipflussdiagramm für die beim Verfahren auszuführenden Schritte; und

Fig. 2 zeigt ein Blockdiagramm zur Realisierung des Verfahrens in einem System, in dem Spracherkennung verwendet wird.

In Fig. 1 ist das erfindungsgemäße Spracherkennungsverfahren klargestellt. Das Verfahren steht nicht in unmittelbarem Zusammenhang mit dem internen, bei der Spracherkennung verwendeten Verfahren der Spracherkennungseinrichtung, sondern unter Verwendung des Verfahrens wird das Erzielen einer Erkennungsauflösung beschleunigt und die Erkennungsgenauigkeit wird verbessert, ohne dass den Eigenschaften der vorliegenden Spracherkennungseinrichtung Aufmerksamkeit zu schenken wäre. Wenn die Spannung in der Einrichtung eingeschaltet wird 1, wird der Speicher rückgesetzt und es wird erwartet, dass eine Äußerung 2 von einem Benutzer erfolgt, wodurch die Spracherkennungseinrichtung Wahrscheinlichkeiten für alle Bezugswörter und als Erkennungsergebnis berechnet 2, und sie das Bezugswort liefert, das die größte Wahrscheinlichkeit besitzt, d.h. dasjenige Bezugswort, das dem vom Benutzer gesprochenen Wort am meisten ähnelt. Wenn die Wahrscheinlichkeit für das Bezugswort einen vorbestimmten Schwellenwert oder den Schwellenwert für die Wahrscheinlichkeiten des wahrscheinlichsten und des zweitwahrscheinlichsten Worts, die im vorliegenden Zusammenhang gemeinsam als Schwellenwerte bei der Spracherkennung bezeichnet werden, nicht überschreitet, wird herausgefunden 3, ob das untersuchte Wort eine Wiederholung des vorangegangenen Worts ist. Wenn eine Wiederholung eines derartigen vorangegangenen Worts nicht zur Debatte steht, wird der Speicher rückgesetzt 4a. Wenn der Benutzer das Wort nicht öfter als einmal gesprochen hat, enthält der Speicher während der ersten Berechnungsrunde nichts, wodurch auch keine neue Wahrscheinlichkeit berechnet wird sondern eine Erkennungsauflösung erzeugt wird 6a, und wobei, wenn keine zuverlässige Erkennung vorgenommen werden kann 6b, die durch die Spracherkennungseinrichtung berechnete Wahrscheinlichkeit in den Speicher eingespeichert wird 7 und auf eine anschließende Äußerung des Benutzers gewartet wird. Wenn dagegen das Wort eine Wiederholung des vorigen Worts ist, wird eine neue Wahrscheinlichkeit berechnet 5, wozu bei der Berechnung der Wahrscheinlichkeit ein im Speicher gespeicherter voran-

gegangener Erkennungsversuch genutzt wird, und auf Grundlage derselben eine Erkennungsauflösung erzeugt 6a, 6b. Wenn die neue Wahrscheinlichkeit dadurch erhalten wird, dass die Berechnungen 5 den Schwellenwert überschreiten, d.h., dass eine zuverlässige Erkennung erfolgen kann 6b, wird der Speicher rückgesetzt 4b und es wird erwartet, dass eine anschließende Äußerung 2 vom Benutzer und ein von der Spracherkennungseinrichtung erhaltenes 2 Erkennungsergebnis auftreten usw. Wenn die neue Wahrscheinlichkeit unter dem Schwellenwert liegt, so dass keine zuverlässige Erkennung erfolgen kann, wird die neue Wahrscheinlichkeit in den Speicher 7 eingespeichert und es wird erwartet, dass eine anschließende Äußerung 2 des Benutzers erfolgt usw. Wenn eine der Funktionen unterbrochen wird, wird der Speicher rückgesetzt, so dass nichts in ihm verbleibt, was eine nach der Unterbrechung zu startende neue Erkennung stören würde. Das erfindungsgemäße Verfahren kann auch so realisiert werden, dass die Erkennungsauflösung 6a, 6b erzeugt wird, bevor herausgefunden wird 3, ob eine Wiederholung des vorangegangenen Worts zur Debatte steht oder nicht. Wenn der von der Spracherkennungseinrichtung für das wiederholte Wort berechnete Wert nun den eingestellten Schwellenwert überschreitet, muss keine derartige Berechnung einer neuen Wahrscheinlichkeit erfolgen, bei der die bei vorangegangenen Erkennungsversuchen berechneten Werte berücksichtigt würden.

Um den Rechenprozess auszuführen, können mehrere Berechnungsabläufe entwickelt werden, bei deren Verwendung eine genauere Wahrscheinlichkeit unter Verwendung der vorangegangenen Wahrscheinlichkeit erzielt werden kann. Jedoch ist die nützlichste Formel die Berechnungsformel für bedingte Wahrscheinlichkeit. Um den bei diesem Verfahren verwendeten Berechnungsablauf zu demonstrieren, wird unten die Verwendung einer Berechnung mit bedingter Wahrscheinlichkeit im einzelnen und in Zusammenhang mit dem erfindungsgemäßen Verfahren beschrieben. Es wird eine Situation untersucht, bei der ein Benutzer als erstes ein Wort A und dann ein Wort B spricht, nachdem er vom System dazu aufgefordert wurde, das Wort zu wiederholen. Eine Spracherkennungseinrichtung berechnet z. B. die folgenden Wahrscheinlichkeiten für die beiden Wörter A und B:

P(A=1) = 0,7 (Wahrscheinlichkeit, dass A "eins" war)
P(A=2) = 0,3 (Wahrscheinlichkeit, dass A "zwei" war)
P(B=1) = 0,8 (Wahrscheinlichkeit, dass B "eins" war)
P(B=1) = 0,2 (Wahrscheinlichkeit, dass B "zwei" war)

Wenn als Schwellenwert für die Erkennungsauflösung 0,9 eingestellt ist,

kann betreffend jede Erkennung keine Erkennungsauflösung erzeugt werden. Wenn bekannt ist, dass der Benutzer beide Male dasselbe Wort sprach, kann die Zuverlässigkeit der Erkennung dadurch erhöht werden, dass zum Berechnen einer neuen Wahrscheinlichkeit die Wahrscheinlichkeit genutzt wird, die hinsichtlich vorangegangener und aktueller Erkennungen durch einen oder mehrerer dieser Vorgänge berechnet wurde. Dies kann z. B. durch eine Berechnung mit bedingter Wahrscheinlichkeit wie folgt erfolgen:

$$P(B=1/A=B) = [P(B=1 \text{ und } A=B) / P(A=B)] =$$
$$= [P(B=1 \text{ und } ((A=1 \text{ und } B=1) \text{ oder } (A=2 \text{ und } B=2)))] / P(A=B)$$
$$= [P((A=1 \text{ und } B=1) \text{ oder } (B=1 \text{ und } A=2 \text{ und } B=2)) / P(A=B)]$$
$$= [P(A=1 \text{ und } B=1) / P((A=1 \text{ und } B=1) \text{ oder } (A=2 \text{ und } B=2)))]$$
$$= [0,7 * 0,8 / 0,7 * 0,8 + 0,3 * 0,2] = 0,56 / 0,62 = 0,903$$

Die obige Berechnung, durch die eine Wahrscheinlichkeit für das Detail berechnet wurde, dass das zweite Wort, d.h. B, "eins" ist, wobei die Bedingung besteht, dass A mit B übereinstimmt, anders gesagt, dass das erste Wort dasselbe wie das zweite Wort ist, führt zu einer neuen Wahrscheinlichkeit, die im vorliegenden Fall den Schwellenwert überschreitet, so dass eine Erkennungsauflösung erzeugt werden kann. Selbst wenn die neue Wahrscheinlichkeit den Schwellenwert nicht überschreitet, ist sie jedoch besser als die durch die Spracherkennungseinrichtung berechnete individuelle Wahrscheinlichkeit, und auf diese Art wird im Speicher eine neue Wahrscheinlichkeit gespeichert und bei der Berechnung einer folgenden, neuen Wahrscheinlichkeit zusammen mit einer folgenden, von der Spracherkennungseinrichtung berechneten Wahrscheinlichkeit verwendet. Es zeigt sich auch, dass der Unterschied zum zweitwahrscheinlichsten Wort zunimmt. Die obige Formel kann dadurch vereinfacht werden, dass nur der Zähler an Stelle des Nenners verwendet wird und mit einer geeigneten Konstanten Y multipliziert wird:

$$P(B=x|A=B) = Y^*P(A=x \text{ und } B=x) = Y^*P(A=x)^*P(B=x)$$

Demgemäß wird die Gesamtwahrscheinlichkeit für jedes Bezugswort r wie folgt erhalten, wenn der Benutzer ein Wort N mal ausspricht:

$$P(r) = Y^*P(r,1)^*P(r,2)^*...^*P(r,N),$$

wobei P(r,1) die erste Äußerung des Bezugsworts r ist, P(r,2) die zweite Äußerung ist und N die letzte Äußerung desselben ist. Beim obigen Beispiel wurde eine Wahrscheinlichkeit für ein gegebenes Bezugswort berechnet. In

Übereinstimmung mit den Schwellenkriterien bei der Spracherkennung nimmt die Differenz zwischen den Wahrscheinlichkeiten zweier Bezugswörter (für das Bezugswort, das von der Spracherkennungseinrichtung die höchste Wahrscheinlichkeit und die zweithöchste Wahrscheinlichkeit erhielt) automatisch zu, weswegen die Erkennungszuverlässigkeit verbessert ist. Es ist einfach, die obigen Rechenverfahren zu verwenden, wenn in der Spracherkennungseinrichtung das HMM-Verfahren verwendet wird, da es in solchen Fällen für jedes Bezugswort die Wahrscheinlichkeit des vom Benutzer gesprochenen Worts berechnet. Wenn das DTW-Verfahren verwendet wird, ist die Berechnung nicht ganz so unkompliziert, da nun für Bezugswörter in der Spracherkennungseinrichtung keine Wahrscheinlichkeit berechnet wird, sondern ein Abstand oder ein Standard dafür, wie weit das gesprochene Wort von jedem Bezugswort entfernt ist.

Daher muss zum Verbessern der Erkennungszuverlässigkeit beim Verfahren, bei dem vorige Wahrscheinlichkeiten genutzt werden, der Standard oder der Abstand als erstes in eine Wahrscheinlichkeit umgewandelt werden. Beim DTW-Verfahren ist es so möglich, mittels einer Zahl D(r,i) zu beschreiben, in welchem Ausmaß jedes Bezugswort r einem gesprochenen Wort innerhalb einer Wiederholungszeit i ähnelt. Hierbei kann eine Wahrscheinlichkeit wie folgt unter Zuhilfenahme einer Funktion f(), z. B. einer nichtlinearen Funktion, aus der Zahl berechnet werden:

$$D(r) = f(D(r,1), D(r,2), \ldots, D(r,N))$$

Alternativ kann ein Schätzwert für die Wahrscheinlichkeit eines Bezugsworts aus dem durch einen DTW-Algorithmus gelieferten Ergebnis mittels eines Schätzwerts g() berechnet werden, wodurch das von der Spracherkennungseinrichtung berechnete Ergebnis in eine Wahrscheinlichkeit umgewandelt werden kann, und die Wahrscheinlichkeit einer i:n-ten Wiederholung eines Bezugsworts r ist nun $P(r,i) = g(D(r,i))$, wobei die Zahl $P(r,i)$ entsprechend dem Verfahren beim Berechnen einer neuen Wahrscheinlichkeit verwendet werden kann, wie oben beschrieben.

In Fig. 2 ist ein Weg zum Realisieren des erfindungsgemäßen Verfahrens in einem Spracherkennungssystem dargestellt. Durch dieses Verfahren kann die Erkennungsgenauigkeit des Spracherkennungssystems verbessert werden, in dem die Spracherkennungseinrichtung 8 Erkennungsergebnisse, d.h. Erkennungswahrscheinlichkeiten, liefert, die an die Verarbeitungseinheit 9 für Erkennungsergebnisse geliefert werden. Jedes Erkennungsergebnis enthält eine

Liste der zu erkennenden Wörter, wobei für jedes eine Wahrscheinlichkeit (oder ein anderer Qualitätsfaktor) berechnet wurde, die beschreibt, in welchem Ausmaß ein vom Benutzer gesprochenes Wort Ähnlichkeit mit jedem Bezugswort hat. Die Bezugswörter können vorab im internen Bezugswörterspeicher der Spracherkennungseinrichtung 8 eingespeichert sein oder die Spracherkennungseinrichtung ist mit der Fähigkeit versehen, vom Benutzer gesprochene Wörter zu "lernen". Jedoch hat dieses Detail dazu, wie und wann Bezugswörter in den Bezugswörterspeicher eingespeichert werden, keine Bedeutung hinsichtlich der Erfindung, und die Spracherkennungseinrichtung 8 muss keinen Bezugswörterspeicher aufweisen. Wenn ein Wort nicht mit ausreichender Zuverlässigkeit erkannt werden kann, fordert die Benutzerkommunikationseinrichtung 11 den Benutzer dazu auf, das Wort zu wiederholen. In einem solchen Fall liefert die Benutzerkommunikationseinrichtung 11 Information an den Verarbeitungsblock 9 für Verarbeitungsergebnisse dahingehend, ob ein Wort vom Benutzer zu wiederholen ist oder nicht. Wenn die Benutzerkommunikationseinrichtung 11 die Verarbeitungseinheit 9 darüber informiert, dass eine Wiederholung eines Worts zu erwarten ist, wird auf die in Verbindung mit dem vorangegangenen Erkennungsversuch gespeicherten Daten aus dem Speicher 11 zugegriffen und für die Bezugswörter werden neue Wahrscheinlichkeiten für die Bezugswörter gemäß der Erfindung auf eine Weise berechnet, die die vorangegangenen Werte berücksichtigt. Wenn keine ausreichend zuverlässige Erkennung, selbst auf Grundlage der neuen Wahrscheinlichkeiten, erfolgen kann, werden diese neuen, genau berechneten Wahrscheinlichkeiten dennoch in den Speicher 10 eingespeichert. Nachdem eine erfolgreiche Erkennung erfolgte, wird der Speicher 10 rückgesetzt. Der Speicher wird auch dann rückgesetzt, wenn Daten von der Benutzerkommunikationseinrichtung dahingehend an den Verarbeitungsblock 9 geliefert werden, dass das nächste eingegebene Wort nicht dasselbe wie das vorige ist. In der Praxis kann das System dergestalt sein, dass der Verarbeitungsblock 9, der Speicher 10 und der Benutzerkommunikationsblock 11 einen Teil desselben Prozessors bilden, d.h., dass sie mittels eines Prozessors realisiert sind. Der Prozessor kann ein solcher sein, der speziell für das Spracherkennungssystem ausgebildet ist, oder es kann der Hauptprozessor für ein Funktelefon sein. Typischerweise verfügt auch die Spracherkennungseinrichtung 9 über einen Signalprozessor.

Unter Zuhilfenahme der Erfindung kann die Spracherkennungsgenauigkeit verbessert werden, obwohl die Grundfunktion der Spracherkennungseinrichtung selbst nicht verbessert ist. Wenn die Erkennungsgenauigkeit verbessert ist, ist die Entscheidungsfindung betreffend Erkennung beschleunigt und es kann

ein benutzerfreundlicheres Freisprechtelefon realisiert werden. Die Erfindung ist nicht auf die Formel des Beispiels, wie in Fig. 1 dargestellt, beschränkt, sondern es können verschiedene Funktionen auch mit anderer Reihenfolge ausgeführt werden.

Angesichts der vorstehenden Beschreibung ist es für den Fachmann ersichtlich, dass innerhalb des Schutzumfangs der durch die beigefügten Ansprüche definierten Erfindung verschiedene Modifizierungen erfolgen können.

Patentansprüche

1. Spracherkennungsvorrichtung mit:

- einer Vergleichseinrichtung zum Vergleichen eines von einem Benutzer gesprochenen ersten Worts mit mindestens einem vorbestimmten Bezugswort;
- einer Berechnungseinrichtung zum Berechnen eines Werts, der der Ähnlichkeit zwischen dem vom Benutzer gesprochenen ersten Wort und dem mindestens einen vorbestimmten Bezugswort entspricht;
- einer Auswähleinrichtung zum Auswählen des Werts, der der größten Wahrscheinlichkeit entspricht;

dadurch gekennzeichnet, dass die Berechnungseinrichtung so ausgebildet ist, dass sie den ausgewählten Wert beim Berechnen eines neuen Werts entsprechend der Ähnlichkeit zwischen einem zweiten vom Benutzer gesprochenen Wort und dem mindestens einen Bezugswort verwendet, wenn der ausgewählte Wert einem vorbestimmten Kriterium genügt.

2. Spracherkennungsvorrichtung nach Anspruch 1 mit mehr als einem vorbestimmten Bezugswort.

3. Spracherkennungsvorrichtung nach Anspruch 1 oder Anspruch 2, bei der das vom Benutzer gesprochene zweite Wort dasselbe wie das vom Benutzer gesprochene erste Wort ist.

4. Spracherkennungsvorrichtung nach einem der vorstehenden Ansprüche, bei der die Berechnungseinrichtung den ausgewählten Wert beim Berechnen eines neuen Werts nur dann verwendet, wenn das vom Benutzer gesprochene zweite Wort dasselbe wie das vom Benutzer gesprochene erste Wort ist.

5. Spracherkennungsvorrichtung nach einem der vorstehenden Ansprüche, bei der das vorbestimmte Kriterium dasjenige ist, dass der ausgewählte Wert kleiner als ein vorbestimmter Schwellenwert ist.

6. Spracherkennungsvorrichtung nach einem der Ansprüche 2 bis 4, bei der das vorbestimmte Kriterium darin besteht, dass die Differenz zwischen dem ausgewählten Wert und einem anderen Wert entsprechend der Ähnlichkeit zwischen dem vom Benutzer gesprochenen ersten Wort und einem anderen Bezugswort kleiner als ein vorbestimmter Schwellenwert ist.

7. Spracherkennungsvorrichtung nach Anspruch 5 oder Anspruch 6, bei der die Wiederholung des vom Benutzer gesprochenen ersten Worts angefordert wird, wenn der ausgewählte Wert dem vorbestimmten Kriterium genügt.

8. Spracherkennungsvorrichtung nach Anspruch 5 oder Anspruch 6, bei der dann, wenn der ausgewählte Wert dem vorbestimmten Kriterium nicht genügt, die Vorrichtung rückgesetzt wird und sie auf eine weitere Äußerung durch den Benutzer wartet.

9. Spracherkennungsvorrichtung nach einem der vorstehenden Ansprüche, bei der der Wert und der neue Wert jeweilige Wahrscheinlichkeiten dafür sind, dass das erste gesprochene Wort und das zweite gesprochene Wort jeweils dem mindestens einen vorbestimmten Bezugswort entsprechen.

10. Spracherkennungsvorrichtung nach Anspruch 9, bei der der Wert und der neue Wert unter Verwendung einer Berechnung mit bedingter Wahrscheinlichkeit berechnet werden.

11. Spracherkennungsvorrichtung nach einem der vorstehenden Ansprüche, mit einer Speichereinrichtung (10) zum Einspeichern des ausgewählten Werts und des neuen Werts.

12. Spracherkennungsverfahren, das folgendes aufweist:
- Vergleichen eines von einem Benutzer gesprochenen ersten Worts mit mindestens einem vorbestimmten Bezugswort;
- Berechnen eines Werts, der der Ähnlichkeit zwischen dem vom Benutzer gesprochenen ersten Wort und dem mindestens einen vorbestimmten Bezugswort entspricht;
- Auswählen des Werts, der der größten Ähnlichkeit entspricht;
dadurch gekennzeichnet, dass der ausgewählte Wert dazu verwendet wird, einen neuen Wert entsprechend der Ähnlichkeit zwischen einem vom Benutzer gesprochenen zweiten Wort und dem mindestens einen Bezugswort zu berechnen, wenn der ausgewählte Wert einem vorbestimmten Kriterium genügt.

13. Verfahren nach Anspruch 12, bei dem mehr als ein vorbestimmtes Bezugswort existiert.

14. Verfahren nach Anspruch 12, bei dem das vom Benutzer gesprochene zweite Wort dasselbe wie das vom Benutzer gesprochene erste Wort ist.

15. Verfahren nach Anspruch 12, bei dem der ausgewählte Wert nur beim Berechnen eines neuen Werts verwendet wird, wenn das vom Benutzer gesprochene zweite Wort dasselbe wie das vom Benutzer gesprochene erste Wort ist.

16. Verfahren nach einem der Ansprüche 12 bis 15, bei dem das vorbestimmte Kriterium dasj nige ist, dass der ausgewählte Wert kleiner als ein vorbestimmter Schwellenwert ist.

17. Verfahren nach einem der Ansprüche 13 bis 15, bei dem das vorbestimmte Kriterium darin besteht, dass die Differenz zwischen dem ausgewählten Wert und einem anderen Wert entsprechend der Ähnlichkeit zwischen dem vom Benutzer gesprochenen ersten Wort und einem anderen Bezugswort kleiner als ein vorbestimmter Schwellenwert ist.

18. Verfahren nach Anspruch 16 oder Anspruch 17, bei dem die Wiederholung des vom Benutzer gesprochenen ersten Worts angefordert wird, wenn der ausgewählte Wert dem vorbestimmten Kriterium genügt.

19. Verfahren nach Anspruch 16 oder Anspruch 17, bei dem dann, wenn der ausgewählte Wert dem vorbestimmten Kriterium nicht genügt, die Vorrichtung rückgesetzt wird und sie auf eine weitere Äußerung durch den Benutzer wartet.

20. Verfahren nach einem der vorstehenden Ansprüche, bei dem der Wert und der neue Wert jeweilige Wahrscheinlichkeiten dafür sind, dass das erste gesprochene Wort und das zweite gesprochene Wort jeweils dem mindestens einen vorbestimmten Bezugswort entsprechen.

21. Verfahren nach Anspruch 20, bei dem der Wert und der neue Wert unter Verwendung einer Berechnung mit bedingter Wahrscheinlichkeit berechnet werden.

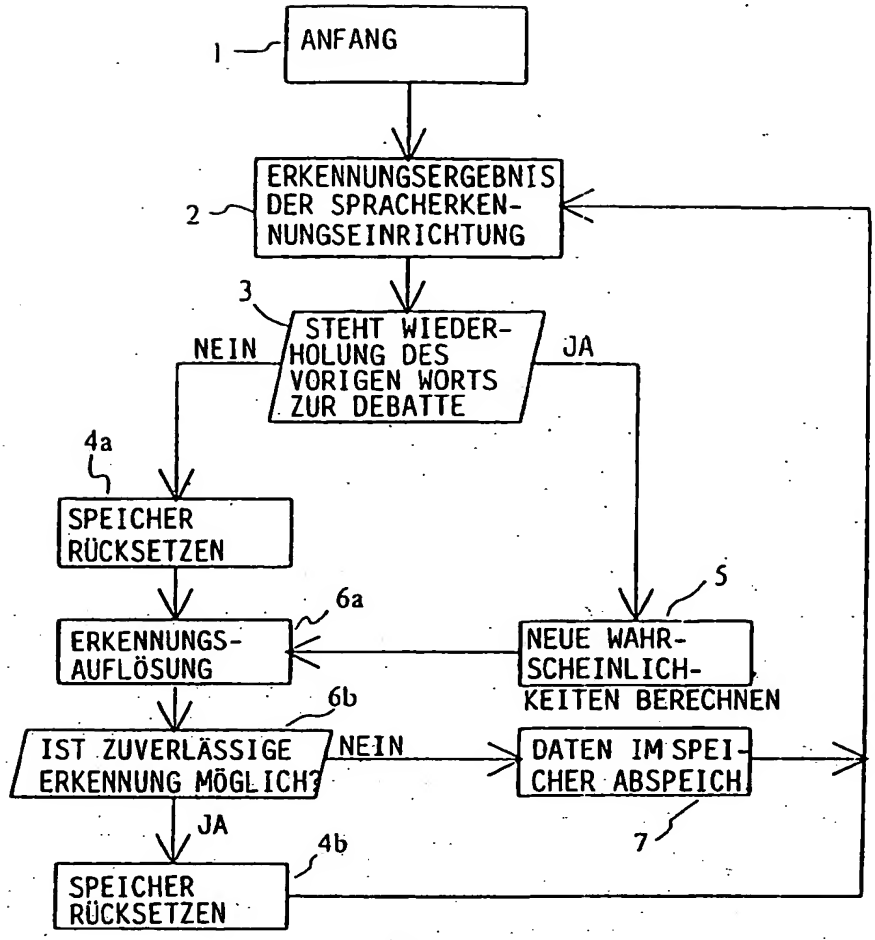
1
ANFANG
2
ERKENNUNGSERGEBNIS DER SPRACHERKENNUNGSEINRICHTUNG
3
STEHT WIEDERHOLUNG DES VORIGEN WORTS ZUR DEBATTE
NEIN
JA
4a
SPEICHER RÜCKSETZEN
6a
ERKENNUNGS-AUFLÖSUNG
5
NEUE WAHRSCHEINLICHKEITEN BERECHNEN
6b
IST ZUVERLÄSSIGE ERKENNUNG MÖGLICH?
NEIN
DATEN IM SPEICHER ABSPEICH
7
JA
4b
SPEICHER RÜCKSETZEN

Fig. 1

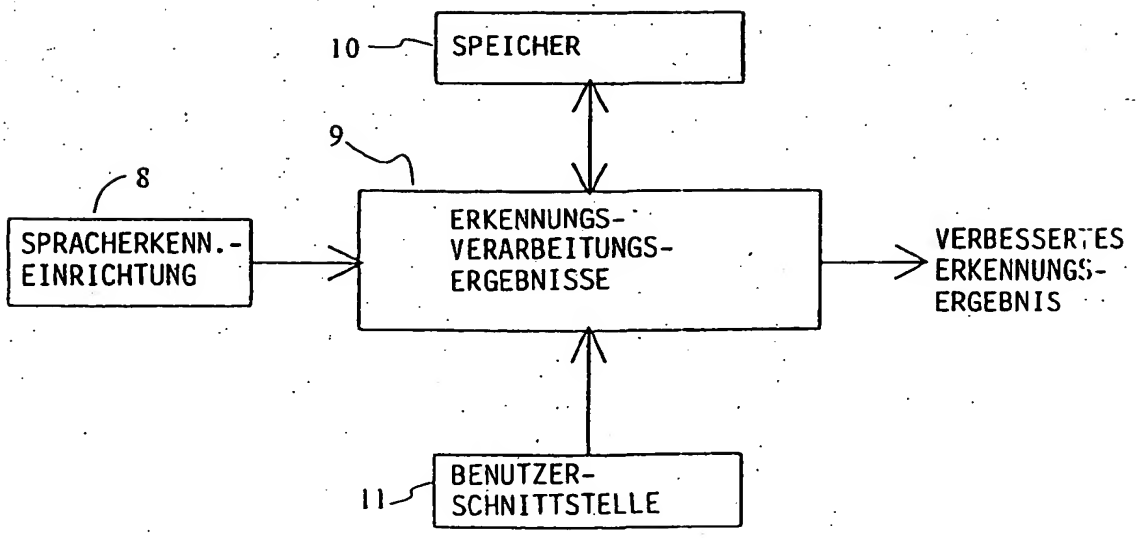
10
SPEICHER
9
8
SPRACHERKENN.-EINRICHTUNG
ERKENNUNGS-VERARBEITUNGS-ERGEBNISSE
VERBESSERTES ERKENNUNGS-ERGEBNIS
11
BENUTZER-SCHNITTSTELLE

Fig. 2